Algebra grammars

RADIM BĚLOHLÁVEK

Abstract. We introduce a new way to describe formal languages. Algebra grammar is a generalization of categorial grammar introduced in [1]. It is shown that both regular and context free languages can be represented by certain types of algebra grammars.

1991 Mathematics Subject Classification:

The aim of this short paper is to show a new way to describe formal languages. Formal languages have been studied by many authors. In [1], the authors proposed the so called *categorial grammars* which have applications in study of natural languages. Categorial grammars are based on the concept of category which can be considered as a syntactical part of a sentence. It is proved that every categorial grammar represents a context free formal language and, on the other hand, every context free formal language can be represented by appropriate categorial grammar.

Let us recall some basic concepts. Let $\langle T(X), F\rangle$ denote the *term algebra* of type F over a countable set X of variables. An *identity* over type F is the expression of the form $p \approx q$ where $p, q \in T(X)$ for some X. A *variety* $\mathcal{V}$ of type F is a non-void class of all algebras of type F which satisfy a given set of identities. The fact that a variety $\mathcal{V}$ satisfies (i.e. each of its members satisfies) the identity I is denoted by $\mathcal{V} \models I$. A *language* L over an alphabet Σ is an arbitrary subset of the set of all strings over Σ, i.e. $L \subseteq \Sigma^*$. Let ϵ denote the empty string.

We are ready to introduce algebra grammars.

Definition. An *algebra grammar* is an ordered 7-tupple

$$\mathcal{AG} = \langle \Sigma, P, \mathcal{V}, F', \circ, s, c\rangle$$

where

- Σ is a finite alphabet.
- P is a finite set of primitive categories
- $\mathcal{V}$ is a variety of algebras of type F, $F \cap P = \emptyset$
- $F' \subseteq F$
- $\circ \in F$ is an associative binary operational symbol, i.e. $\mathcal{V} \models a\circ(b\circ c) \approx (a\circ b)\circ c$
- s is an element of $P \cup F_0$, where F_0 is the set of all nullary operational symbols of F.

- c is a function which assigns to each $a \in \Sigma$ a finite subset of the support of $\langle T(P), F' \rangle$.

Elements of the support of $\langle T(P), F' \rangle$ are called categories.

Definition. A language represented by an algebra grammar $\mathcal{AG} = \langle \Sigma, P, \mathcal{V}, F', \circ, s, c \rangle$ is the set

$$L(\mathcal{AG}) = \{a_1 a_2 \dots a_n \in \Sigma^*; \exists p_i \in c(a_i), i = 1, \dots, n \ : \ \mathcal{V} \models p_1 \circ p_2 \circ \dots \circ p_n \approx s\}.$$

In other words, $a_1 a_2 \dots a_n \in L(\mathcal{V})$ if and only if we can find appropriate categories p_i (assigned to a_i by the function c) such that we can reduce the product of p_i's to the category s by the "rules of computation" of variety $\mathcal{V}$.

Example. Let $\mathcal{V}$ be the variety of all groups, $F = \{\circ, ^{-1}, e\}$, $F' = \{^{-1}, e\}$, $\Sigma = \{0, 1\}$, $s = e$, $P = \{x\}$, $c(0) = x$, $c(1) = x^{-1}$. Then $L(\mathcal{AG})$ is the set of all strings over $\{0, 1\}$ which have the same number of 0's and 1's.

Proposition. *Let $\mathcal{AG}_1 = \langle \Sigma, P, \mathcal{V}_1, F', \circ, s, c_1 \rangle$, $\mathcal{AG}_2 = \langle \Sigma, P, \mathcal{V}_2, F', \circ, s, c_2 \rangle$ be two algebra grammars, $\mathcal{V}_1 \subseteq \mathcal{V}_2$, $c_1(a) \supseteq c_2(a)$ for each $a \in \Sigma$. Then $L(\mathcal{AG}_1) \supseteq L(\mathcal{AG}_2)$.*

PROOF: The proof follows immediately from the fact that if $\mathcal{V}_1 \subseteq \mathcal{V}_2$ then $\mathcal{V}_2 \models I$ implies $\mathcal{V}_1 \models I$ for each identity I. □

Definition. A class $\mathcal{L}$ of languages is representable by $(\mathcal{V}, F')$ where $\mathcal{V}$ is a variety of type F, $F' \subseteq F$, if the following assertions are equivalent :

- $L \in \mathcal{L}$, $\epsilon \notin L$
- $L = L(\mathcal{AG})$ for some algebra grammar $\mathcal{AG} = \langle \Sigma, P, \mathcal{V}, F', \circ, s, c \rangle$.

Let $F_r = \{\rhd, \circ, s, t\}$ where both $\rhd$ and $\circ$ are binary and s, t are nullary. Let $\mathcal{REG}$ be the variety of all algebras of type F_r satisfying the following indentities

$$\begin{aligned} x \circ (y \circ z) &\approx (x \circ y) \circ z \\ (x \rhd y) \circ (y \rhd z) &\approx (x \rhd z) \\ s \rhd t &\approx s \,. \end{aligned}$$

Theorem 1. *The class of all regular languages is representable by $(\mathcal{REG}, \{\rhd, s, t\})$.*

PROOF: (1) Let L be a regular language represented by the regular grammar $G = (N, \Sigma, R, S)$, $\epsilon \notin L$. Rewriting rules of G are of the form $A \to aB$ or $A \to a$ where $A, B \in N$ and $a \in \Sigma$. Consider the algebra grammar $\mathcal{AG} = \langle \Sigma, N, \mathcal{REG}, \{\rhd, s, t\}, \circ, s, c \rangle$ where

$$\begin{aligned} c(a) \;=\; & \{s \rhd A; S \to aA \in R\} \cup \{A \rhd B;\ A \to aB \in R\} \cup \\ & \{A \rhd t;\ A \to a \in R\} \cup \{s \rhd t;\ S \to a \in R\} \end{aligned}$$

Let $a_1a_2\ldots a_n \in L(G)$. For $n = 1$, we have $S \to a_1 \in R$ thus $s \triangleright t \in c(a_1)$. Because of $\mathcal{V} \models (s \triangleright t) \approx s$, we have $a_1 \in L(\mathcal{AG})$. If $n\rangle 1$, there are $S \to a_1A_1$, $A_1 \to a_2A_2,\ldots,A_{n-1} \to a_n \in R$. But then $s \triangleright A_1 \in c(a_1)$, $A_1 \triangleright A_2 \in c(a_2),\ldots$, $A_{n-1} \triangleright t \in c(a_n)$. We can easily see that the identity

$$(s \triangleright A_1) \circ (A_1 \triangleright A_2) \circ \ldots \circ (A_{n-1} \triangleright t) \approx s$$

holds in $\mathcal{REG}$. Thus $a_1a_2\ldots a_n \in L(\mathcal{AG})$, $L(G) \subseteq L(\mathcal{AG})$.
Let $a_1a_2\ldots a_n \in L(\mathcal{AG})$. For $n = 1$, there must be $s \triangleright t \in c(a_1)$ thus $S \to a_1 \in R$ and $a_1 \in L(G)$. For $n\rangle 1$, there must be $s \triangleright A_1 \in c(a_1)$, $A_1 \triangleright A_2 \in c(a_2),\ldots$, $A_{n-1} \triangleright t \in c(a_n)$ which implies the existence of the rules $S \to a_1A_1$, $A_1 \to a_2A_2,\ldots$, $A_{n-1} \to a_n$, so $a_1a_2\ldots a_n \in L(G)$, $L(\mathcal{AG}) \subseteq L(G)$. We have proved that L is represented by $\mathcal{AG}$.
(2) Let $\mathcal{AG} = \langle \Sigma, P, \mathcal{REG}, \{\triangleright, s, t\}, \circ, s, c\rangle$ be an algebra grammar. The case $L(\mathcal{AG}) = \emptyset$ is trivial. Let $N' = \{p;\, p \triangleright q \in c(a) \text{ or } q \triangleright p \in c(a) \text{ for some } a \in \Sigma\} \cup \{s\}$. Evidently, the relation $\mathcal{E}$ on N' defined by $\langle p, q\rangle \in \mathcal{E}$ if and only if $\mathcal{REG} \models p \approx q$ is an equivalence. Put $N = N'/\mathcal{E}$ and for $p \in N'$ denote $[p]$ the equivalence-class containing p. Consider the regular grammar $G = (N, \Sigma, R, [s])$ where

$$\begin{aligned} R \;=\; & \{[p] \to a[q];\, p \triangleright q \in c(a)\} \cup \{[p] \to a;\, p \triangleright t \in c(a)\} \cup \\ & \{[s] \to a;\ \text{there is } p \in c(a) \text{ such that } \mathcal{REG} \models p \approx s\} \cup \\ & \{[s] \to a[t];\, s \in c(a)\}. \end{aligned}$$

Let $a_1a_2\ldots a_n \in L(G)$. Denote $r_0 = s$. Then there are $[r_0] \to a_1[r_1]$, $[r_1] \to a_2[r_2],\ldots [r_{n-1}] \to a_n \in R$. If $[r_{i-1}] \to a_i[r_i] \neq [s] \to a_i[t]$ for some $i = 1,\ldots,n-1$ then there is $p_i = r'_{i-1} \triangleright r'_i \in c(a_i)$ such that the identities $r_{i-1} \approx r'_{i-1}$ and $r_i \approx r'_i$ hold in $\mathcal{REG}$. For $[r_{i-1}] \to a_i[r_i] = [s] \to a_i[t]$ there is $p_i \in c(a_i)$ such that $\mathcal{REG} \models p_i \approx s$. If $[r_{n-1}] \neq [s]$ then there is $p_n = r'_{n-1} \triangleright t \in c(a_n)$ such that $\mathcal{REG} \models r'_{n-1} \approx r_{n-1}$. If $[r_{n-1}] = [s]$ then there is $p_n \in c(a_n)$ such that $\mathcal{REG} \models p_n \approx s$. Denote $p'_i = s \triangleright t$ if $\mathcal{REG} \models p_n \approx s$ and $p'_i = p_i$ otherwise, $i = 1,\ldots,n$. Clearly $p'_1 \circ p'_2 \circ \ldots \circ p'_n \approx s$ and $p_i \approx p'_i$ for $i = 1,\ldots,n$ hold in $\mathcal{REG}$ which proves $a_1a_2\ldots a_n \in L(\mathcal{AG})$.

Let $a_1a_2\ldots a_n \in L(\mathcal{AG})$. There are $p_i \in c(a_i)$, $i = 1,\ldots,n$, such that $p_1 \circ \ldots \circ p_n \approx s$ holds in $\mathcal{REG}$. Denote $p'_i = p_i$ for $p_i = p \triangleright q$ for some p, q, $p'_i = s \triangleright t$ for $p_i = s$. Then $p'_1 \circ \ldots \circ p'_n \approx p_1 \circ \ldots \circ p_n$ holds in $\mathcal{REG}$. Denote $p'_i = r_i \triangleright r'_i$. Because of $p'_1 \circ \ldots \circ p'_n \approx s$, the identities $r_1 \approx s$, $r'_i \approx r_{i+1}$ (for $i = 1, ..., n-1$), $r'_n \approx t$ hold in $\mathcal{REG}$. By definition of R, there are $[r_1] \to a_1[r'_1],\ldots,[r_n] \to a_n \in R$. But $[r_1] = [s]$ and $[r'_i] = [r_{i+1}]$ for $i = 1,\ldots,n-1$, thus $a_1a_2\ldots a_n \in L(G)$. We have proved $L(\mathcal{AG}) = L(G)$. The proof is complete. □

Let $F_{cf} = \{/, \backslash, \circ\}$ where each of F_{cf} is binary. Let $\mathcal{CF}$ be the variety of all algebras of type F_{cf} satisfying

$$\begin{aligned} x \circ (y \circ z) &\approx (x \circ y) \circ z \\ x \circ (x \backslash y) &\approx y \\ (y \circ x)/x &\approx y\,. \end{aligned}$$

Theorem 2. *The class of all context free languages is representable by* $(\mathcal{CF}, \{/, \backslash\})$.

PROOF: It is evident that a language L is representable by some algebra grammar $\langle \Sigma, P, \mathcal{CF}, \{/, \backslash\}, \circ, s, c\rangle$, $s \in P$, if and only if it is determined by the bidirectional categorial grammar (see [1]) (Σ, Cat, s, c) where Cat is the support of $\langle T(P), \{/, \backslash\}\rangle$. Following [1], a language L is context free if and only if it is determined by some bidirectional category grammar which completes our proof. □

References

[1] Bar-Hillel, Y., Gaifman, C., Shamir, E., *On categorial and phrase-structure grammars*, Bull. Res. Counc. of Israel, Vol. 9F, 1960.

[2] Burris, S., Sankappanavar, H. P., *A course to universal algebra*, Springer-Verlag, New York, 1981.

Address: Dept. of Computer Science
Technical University of Ostrava
tř. 17. listopadu
708 33 Ostrava-Poruba
Czech Republic